SenseDefs : a multilingual corpus of semantically annotated textual definitions

Camacho-Collados, Jose

2019

Camacho-Collados , J , Delli Bovi , C , Raganato , A & Navigli , R 2019 , ' SenseDefs : a multilingual corpus of semantically annotated textual definitions ' , Language Resources and þÿ E v a l u a t i o n , v o l . 5 3 , n o . 2 , p p . 2 5 1 2 7 8 . h t t p s : / / d o i . o r g / 1 0 . 1 0 0 7 / s 1 0

http://hdl.handle.net/10138/304223
https://doi.org/10.1007/s10579-018-9421-3

Lang Resources & Evaluation
https://doi.org/10.1007/s10579-018-9421-3

ORIGINAL PAPER

SENSEDEFS: a multilingual corpus of semantically annotated textual definitions

Exploiting multiple languages and resources jointly for high-quality Word Sense Disambiguation and Entity Linking

Jose Camacho-Collados[1] · Claudio Delli Bovi[2] · Alessandro Raganato[3] · Roberto Navigli[4]

Abstract Definitional knowledge has proved to be essential in various Natural Language Processing tasks and applications, especially when information at the level of word senses is exploited. However, the few sense-annotated corpora of textual definitions available to date are of limited size: this is mainly due to the expensive and time-consuming process of annotating a wide variety of word senses and entity mentions at a reasonably high scale. In this paper we present SENSEDEFS, a large-scale high-quality corpus of disambiguated definitions (or *glosses*) in multiple languages, comprising sense annotations of both concepts and named entities from a wide-coverage unified sense inventory. Our approach for the construction and disambiguation of this corpus builds upon the structure of a large multilingual semantic network and a state-of-the-art disambiguation system: first, we gather

The work of Jose Camacho-Collados, Claudio Delli Bovi and Alessandro Raganato was mainly done at Sapienza University of Rome.
http://lcl.uniroma1.it/sensedefs.

✉ Jose Camacho-Collados
camachocolladosj@cardiff.ac.uk

Claudio Delli Bovi
boviclau@amazon.com

Alessandro Raganato
alessandro.raganato@helsinki.fi

Roberto Navigli
navigli@di.uniroma1.it

[1] Cardiff University, Cardiff, UK

[2] Amazon.com, Inc., Turin, Italy

[3] University of Helsinki, Helsinki, Finland

[4] Sapienza University of Rome, Rome, Italy

Published online: 23 July 2018

complementary information of equivalent definitions across different languages to provide context for disambiguation; then we refine the disambiguation output with a distributional approach based on semantic similarity. As a result, we obtain a multilingual corpus of textual definitions featuring over 38 million definitions in 263 languages, and we publicly release it to the research community. We assess the quality of SenseDefs's sense annotations both intrinsically and extrinsically on Open Information Extraction and Sense Clustering tasks.

Keywords Textual definitions · Glosses · Word Sense Disambiguation · Entity linking · Multilinguality · Lexical resources

1 Introduction

In addition to lexicography, where their use is of paramount importance, textual definitions drawn from dictionaries or encyclopedias have been widely used in various Natural Language Processing (NLP) tasks and applications. Definitional knowledge is effective inasmuch as it conveys the crucial semantic information and the distinguishing features of a given subject (*definiendum*): this means that, on the one hand, a definition often provides a fair amount of discriminative power that can be leveraged to automatically represent and disambiguate the definiendum; on the other, definitions are usually concise and encode "dense", virtually noise-free information that can be best exploited with knowledge acquisition techniques. To date, some of the areas where the use of definitional knowledge has proved to be key in achieving state-of-the-art results are Word Sense Disambiguation (Lesk 1986; Banerjee and Pedersen 2002; Navigli and Velardi 2005; Agirre and Soroa 2009; Faralli and Navigli 2012; Fernandez-Ordonez et al. 2012; Chen et al. 2014; Basile et al. 2014; Camacho-Collados et al. 2015b), Taxonomy and Ontology Learning (Velardi et al. 2013; Flati et al. 2016; Espinosa-Anke et al. 2016b), Information Extraction (Richardson et al. 1998; Delli Bovi et al. 2015), Plagiarism Detection (Franco-Salvador et al. 2016), and Question Answering (Hill et al. 2015).

In fact, textual definitions are today widely available in knowledge resources of various kinds, ranging from lexicons and dictionaries, such as WordNet (Miller et al. 1990) or Wiktionary, to encyclopedic knowledge bases, such as Wikidata (see Sect. 2 for a brief overview). Interestingly enough, sources of definitional knowledge also include Wikipedia: despite its purely encyclopedic nature, and although the format of a Wikipedia article does not include an explicit gloss or definition, the first sentence of each article is generally regarded as the definition of its subject.

Irrespective of the nature of the knowledge source, an accurate semantic analysis of textual definitions is made difficult by the short and concise nature of definitional text, a crucial issue for automatic disambiguation systems that rely heavily on local context. Furthermore, the majority of approaches making use of definitions are restricted to corpora where each concept or entity is associated with a single definition; instead, definitions coming from different resources are often complementary and might give different perspectives on the definiendum. Moreover,

equivalent definitions of the same concept or entity may vary substantially according to the language, and be more precise or self-explanatory in some languages than others. In fact, the way a certain concept or entity is defined in a given language is sometimes strictly connected to the social, cultural and historical background associated with that language, a phenomenon that also affects the lexical ambiguity of the definition itself. This difference in the degree of ambiguity when moving across languages is especially valuable in the context of disambiguation (Navigli 2012), as highly ambiguous terms in one language may become less ambiguous (or even unambiguous) in other languages.

The fundamental idea of this paper is to bring together definitions coming from different resources and different languages, and disambiguate them by exploiting their cross-lingual and cross-resource complementarities. Our goal is to obtain a large-scale high-quality corpus of sense-annotated textual definitions, constructed using a single multilingual disambiguation model. While language- and resource-specific techniques can certainly be used for disambiguation, they would not be scalable for our goal: the number of models required would add up to the order of hundreds, and there would also be the need for large amounts of sense-annotated data for each language and resource, leading to the so-called knowledge acquisition bottleneck (Gale et al. 1992).

A key step in achieving our goal is to leverage BabelNet (Navigli and Ponzetto 2012), a multilingual lexicalized semantic network obtained from the automatic integration of lexicographic and encyclopedic resources. Thanks to its wide coverage of both lexicographic and encyclopedic terms, BabelNet provides a very large sense inventory for disambiguation, and at the same time a vast and comprehensive target corpus of textual definitions. In fact, as it is a merger of various different resources, BabelNet provides a heterogeneous set of over 35 million definitions for over 250 languages from WordNet, Wikipedia, Wiktionary, Wikidata and OmegaWiki. To the best of our knowledge, this set constitutes the largest available corpus of definitional text.

In this paper we present SENSEDEFS, a large multilingual corpus of sense-annotated glosses. This resource is based on our approach presented in Camacho-Collados et al. (2016a). In the present paper we provide the following main contributions with respect to our prior study:

1. We provide an exhaustive background of all the resources used in this work (Sect. 2), as well as an extended step-wise description of our disambiguation pipeline (Sect. 3).
2. We present an overview of the resource, including relevant statistics about its extension and language coverage (Sect. 4.1).
3. In addition to the previous XML format, we release the resource in NIF format, making it compatible with Semantic Web technologies (Sect. 4.2.2).
4. We carry out an extensive manual evaluation of the resource intrinsically for four different languages: English, French, Italian and Spanish (Sect. 5.1.1).
5. We additionally provide a large-scale automatic evaluation on the English WordNet glosses (Sect. 5.1.2).

The remainder of the paper is organized as follows: Sect. 2 provides a brief overview on the individual semantic resources from which the textual definitions inside SENSEDEFS are drawn, and then gives some background information on BabelNet and Babelfy; Sect. 3 describes our disambiguation strategy and details each stage of our pipeline; the final resource obtained as a result, SENSEDEFS, is presented in Sect. 4, while Sect. 5 describes our experimental evaluation; in Sect. 6 we review some related work in the field and draw our conclusions and perspectives on future work in Sect. 7.

2 Background

In this section we provide some background information about the main resources and tools used in this study, namely BabelNet and all its integrated resources (Sect. 2.1), Babelfy (Sect. 2.2) and NASARI (Sect. 2.3).

2.1 BabelNet

BabelNet (Navigli and Ponzetto 2012) is a large-scale, multilingual encyclopedic dictionary (i.e. a resource where both lexicographic and encyclopedic knowledge is available in multiple languages) and at the same time a semantic network where concepts and entities are interconnected with several million semantic relations. Each concept or entity inside BabelNet is associated with a synonym set (*synset*), comprising lexicalizations of that concept or entity in a variety of languages. Originally designed as the seamless integration of WordNet and Wikipedia, BabelNet[1] is the largest resource of its kind, with 13 million synsets, 380 million semantic relations and 271 languages covered.[2] For our purposes, not only does BabelNet represent the largest sense inventory available for disambiguation and entity linking, its internal structure, based on inter-resource mappings, enables us to collect all the definitional knowledge associated with a given definiendum inside the various individual resources and for any available languages. This is a crucial step for context-rich disambiguation (Sect. 3.2). In the following we describe the resources from which the definitions are extracted: WordNet (Sect. 2.1.1), Wikipedia (Sect. 2.1.2), Wikidata (Sect. 2.1.3), Wiktionary and OmegaWiki (Sect. 2.1.4).

2.1.1 WordNet

The Princeton *WordNet* of English (Miller et al. 1990) is by far the most widely used computational lexicon in Natural Language Processing. It is manually curated by expert lexicographers and organized as a semantic network, where concepts are connected via lexico-semantic relations. Its internal structure based on synsets constitutes the backbone of BabelNet (see Sect. 2.1). Similarly to traditional

[1] http://babelnet.org.

[2] These figures correspond to its 3.0 release version, which is the version used in this work.

dictionaries, WordNet provides a textual definition (*gloss*), as well as small usage examples for each synset. Being hand-crafted by expert annotators, definitional knowledge from WordNet is among the most accurate available and includes non-nominal parts of speech rarely covered by other resources (e.g. adjectives and adverbs). However, given its considerably smaller scale, WordNet provides less than 1% of the overall number of definitions in BabelNet.

2.1.2 Wikipedia

Wikipedia[3] is the largest and most popular collaborative multilingual encyclopedia of world and linguistic knowledge. It features articles in over 250 languages, partially structured with hyperlink connections and categories, and today represents an extraordinary resource for innumerable tasks in Natural Language Processing (Cucerzan 2007; Gabrilovich and Markovitch 2007; Wu and Weld 2010; Chen et al. 2017). As already mentioned in Sect. 1, Wikipedia articles do not provide explicit glosses or definitions, however, according to the style guidelines of Wikipedia[4], an article should begin with a short declarative sentence defining what (or who) the subject is and why it is notable. Following previous literature, we also consider the first sentence of a Wikipedia article as a valid definition of the corresponding concept or entity. Furthermore, text snippets drawn from the associated disambiguation pages can also be regarded as definitions.[5] Due to its focus on encyclopedic knowledge, Wikipedia contains almost exclusively nominal senses (such as named entities or specialized concepts); however, compared to lexicographic resources like WordNet (Sect. 2.1.1), definitions drawn from Wikipedia constitute by far the largest individual contribution to SENSEDEFS ($> 77\%$ of the total).

2.1.3 Wikidata

Wikidata (Vrandečić 2012) is a project operated directly by the Wikimedia Foundation. Wikidata's goal is to turn Wikipedia into a fully structured resource, thereby providing a common source of data that can be used by other Wikimedia projects. It is designed as a document-oriented semantic database based on *items*, each representing a concept or an entity and associated with a unique identifier. Knowledge is encoded with *statements* in the form of property-value pairs, among which definitions (*descriptions*) are also included. Wikidata is the second largest individual contribution to SENSEDEFS (more than 8 million items and $\simeq 22\%$ of the total), even though, given its strictly computational nature, it often provides minimal definition phrases containing only the superclass of the definiendum.

[3] https://www.wikipedia.org.

[4] https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style.

[5] The release format of SENSEDEFS (Sect. 4.2) specifies two distinct attribute values for definitions extracted from the first sentence of Wikipedia articles (`WIKI`) and definitions extracted from disambiguation pages (`WIKIDIS`).

2.1.4 Wiktionary and OmegaWiki

Beyond WordNet, Wikipedia and Wikidata, the remaining definitions ($\simeq 1\%$ of the total) are provided by two collaborative multilingual dictionaries: Wiktionary and OmegaWiki. *Wiktionary*[6] is a Wikimedia project designed to represent lexicographic knowledge that would not be well suited for an encyclopedia (e.g. verbal and adverbial senses). It is available for over 500 languages typically with a very high coverage, including domain-specific terms and descriptions that are not found in WordNet. Similar to Wiktionary, *OmegaWiki*[7] is a large multilingual dictionary based on a relational database, designed with the aim of unifying the various language-specific Wiktionaries into a unified lexical repository.

2.2 Babelfy

Babelfy (Moro et al. 2014) is a graph-based approach to joint multilingual Word Sense Disambiguation (WSD) and Entity Linking based on a loose identification of candidate meanings, and on a densest-subgraph algorithm to select high-coherence semantic interpretations. Unlike supervised WSD approaches that rely heavily on sense-annotated training data to learn a disambiguation model for each target word (*word expert*), Babelfy's strategy does not require training word-specific models from sense-annotated data, but rather leverages an underlying semantic network to connect all the concepts and entities in its sense inventory. These connections are then used to build *semantic signatures* for each concept and entity, using random walks with restart. With state-of-the-art performances reported on various Word Sense Disambiguation and Entity Linking benchmarks, Babelfy is arguably the optimal choice given the wide-coverage sense inventory of BabelNet and our multilingual disambiguation setting. In our pipeline we crucially leverage Babelfy's coverage and flexibility (Sect. 3.2); being completely unsupervised and language-independent, the algorithm can easily be applied to any language for which lexicalizations are available inside the underlying semantic network. As a result, Babelfy can handle mixed text in which multiple languages are used at the same time, or even work without being supplied with information as to which languages the input text contains (*language-agnostic* setting).

2.3 NASARI

Nasari (Camacho-Collados et al. 2016b) is a vectorial representation of concepts and entities from the BabelNet sense inventory. Nasari leverages structural properties from BabelNet, encyclopedic knowledge from Wikipedia and word embeddings trained on large corpora. Given a BabelNet synset, its Nasari representation is computed by first gathering a relevant sub-corpus of contextual information from Wikipedia, exploiting both the Wikipedia inter-link structure and the BabelNet taxonomy. All content words in this sub-corpus are then tokenized,

[6] https://www.wiktionary.org.

[7] http://www.omegawiki.org.

lemmatized and weighted using *lexical specificity* (Lafon 1980), a statistical measure based on the hypergeometric distribution that measures the relevance of a word in a given sub-corpus.[8] Finally, the sub-corpus is turned into a vector using three different techniques that give rise to three different types of representation: *lexical*, *unified*, and *embedded*. In this paper we rely on the latter type (NASARI-*embed*).[9] The word embeddings used for NASARI-embed are the pre-trained vectors of Word2Vec (Mikolov et al. 2013), trained on the Google News corpus. These 300-dimensional word embeddings are injected into the NASARI embedded representation via a weighted average, where the weights are given by lexical specificity. The resulting vector is still defined at the sense level, but lies in the same semantic space as word embeddings, thus enabling a direct comparison between words and synsets. In this work we use NASARI for refining and improving the sense annotations using semantic similarity (Sect. 3.3). NASARI has proved to be effective in various NLP tasks, including not only semantic similarity and WSD (Shalaby and Zadrozny 2015; Camacho-Collados et al. 2016b; Tripodi and Pelillo 2017), but also sense clustering (see Sect. 5.2.2), knowledge-base construction and alignment (Lieto et al. 2016; Espinosa-Anke et al. 2016a; Camacho-Collados and Navigli 2017; Cocos et al. 2017), object recognition (Young et al. 2016) and text classification (Pilehvar et al. 2017).

3 Methodology

In this section we describe our methodology for disambiguating the target corpus of textual definitions that will go to make up SENSEDEFS. Our goal is to disambiguate textual definitions, as provided by the various lexical resources integrated into BabelNet (cfr. Sect. 2.1), and to obtain as many sense annotations as possible, while at the same time retaining high disambiguation accuracy across languages. To this end, we perform a joint disambiguation of both concepts and entities in three successive stages, using BabelNet as reference sense inventory. Together with a unified sense inventory, BabelNet also provides inter-resource mappings that can be exploited to directly convert and utilize the sense annotations obtained within each individual resource (e.g. WordNet, Wikipedia, Wikidata). Our disambiguation strategy is based on three steps: (1) for a given concept or entity, we first gather all its available definitions, drawn from different resources and in different languages, and construct a multilingual sub-corpus of definitional knowledge (Sect. 3.1); (2) we then perform a first high-coverage disambiguation step on this sub-corpus (Sect. 3.2); and, finally, (3) we refine the disambiguation output at the previous step using a procedure based on distributional semantic similarity (Sect. 3.3).

[8] Lexical specificity has been shown to outperform *tf-idf* as a vector weighting scheme (Camacho-Collados et al. 2015a).

[9] We use NASARI-embed version 3.0, available at lcl.uniroma1.it/nasari.

Fig. 1 Some of the definitions, drawn from different resources and languages, associated with the concept of *castling* in chess through our context enrichment procedure

3.1 Step 1: Harvesting textual definitions in multiple languages and resources

As highlighted in Sect. 1, definitional knowledge is not easy to analyze automatically at the sense level. Since many definitions are short and concise, the lack of sufficient and/or meaningful context might negatively affect the performance of an off-the-shelf disambiguation system that works at the sentence level (i.e. targeting individual definitions one by one). In light of this, we leverage the inter-resource and inter-language mappings provided by BabelNet to combine multiple definitions (drawn from different resources and in different languages) of the same concept or entity; in this way, we can associate a much richer context with each target definition, and enable high-quality disambiguation.

As an example,[10] consider the following definition of *castling* in chess as provided by WordNet:

Interchanging the positions of the king and a rook. (1)

The context in this example is limited and it might not be obvious for an automatic disambiguation system that the concept being defined relates to *chess*: for instance, an alternative definition of *castling* where the game of *chess* is explicitly mentioned would definitely help the disambiguation process. Following this idea, given a BabelNet synset, we carry out a *context enrichment* procedure by collecting all the definitions of this synset in every available language and resource, and gathering them together into a single multilingual text. Figure 1 gives a pictorial representation of this harvesting process for the concept of *castling* introduced in Example 1.

[10] This definition will be used as a running example throughout this section.

3.2 Step 2: Context-rich disambiguation

Once a multilingual text is gathered for a given concept or entity, an initial preprocessing step on all definitions is performed. The preprocessing consists of tokenization, part-of-speech (PoS) tagging and lemmatization for a subset of languages:

- *Tokenization* We use the tokenization system available from the polyglot project[11] for 165 languages.
- *Part-of-speech tagging* We train the Stanford tagger (Toutanova et al. 2003), for 30 languages using the available training data from the Universal Dependencies project[12] (Nivre et al. 2016).
- *Lemmatization* We lemmatize all content words (i.e. nouns, verbs and adjectives) using BABELMORPH[13], an open-source API based on Wiktionary and designed to retrieve the morphology of content words (i.e., nouns, verbs and adjectives) for several languages.

Then, we employ Babelfy (Moro et al. 2014) (see Sect. 2.2) to disambiguate with high coverage all content words in all the available languages at once. Our methodology is based on the fact that knowledge-based disambiguation systems like Babelfy work better with richer context, even when they use no supervision. In fact, at disambiguation time, Babelfy considers the content words across the target text in order to construct an associated semantic graph, whose richness in terms of nodes and edges strictly depends on the number of content words. When provided solely with the English WordNet definition of (1), Babelfy disambiguates *rook* incorrectly as "rookie, inexperienced youth". However, as additional definitions from other resources and languages are included, Babelfy exploits the added context to construct a richer semantic graph, and disambiguates *rook* with its correct chess-related sense. This approach is particularly advantageous for languages with low resources, where standard disambiguation techniques have not yet proven to be effective, due to the lack of sufficient sense-annotated data. As a result of this disambiguation step, we obtain a fully disambiguated corpus of definitions, which is later refined by means of distributional semantic similarity. In the following section we explain how this refinement is carried out.

3.3 Step 3: Disambiguation refinement based on distributional similarity

As output of the previous disambiguation step, we obtained a set D of *disambiguated instances*. These disambiguated instances consist of unambiguous senses from the BabelNet sense inventory, each associated with a confidence score (*Babelfy score* henceforth). However, when the Babelfy score goes below 0.7, a back-off strategy based on the *Most Common Sense* (MCS) is activated by default

[11] http://polyglot.readthedocs.org/en/latest/Tokenization.html.

[12] We used version 1.2, available at https://universaldependencies.github.io/docs.

[13] https://github.com/raganato/BabelMorph.

for that instance. In fact, Babelfy has been shown to be heavily biased towards the MCS (Raganato et al. 2017). At this stage, our task is to reduce this bias by correcting or discarding these low-confidence instances using semantic similarity.

First of all, for each disambiguated instance[14] $d \in D$ we compute a *coherence score* C_d. The coherence score is computed as the number of semantic connections from the BabelNet synset d to any other disambiguated instance in D inside the BabelNet semantic network, divided by the total number of disambiguated instances:

$$C_d = \frac{|\text{Disambiguated instances connected to } d|}{|\text{Disambiguated instances}| - 1} \tag{2}$$

We empirically set a coherence score threshold to 0.125 (i.e. one semantic connection out of eight disambiguated instances). Let L be the set of disambiguated instances below both the Babelfy score and the coherence score thresholds (namely the low-confidence annotations). In order to refine the disambiguated instances in L, we use NASARI (Camacho-Collados et al. 2016b). NASARI provides embedded vector representations for over four million BabelNet synsets which were constructed by exploiting the complementary knowledge of Wikipedia, WordNet and text corpora (see Sect. 2.3). We consider those instances in L for which a NASARI> vector can be retrieved (virtually all noun instances), and compute an additional score (*NASARI score*). First, we calculate the centroid μ of all the NASARI vectors for instances in $D \setminus L$. This centroid represents the vector of maximum coherence, as it corresponds to the point in the vector space which is closer to all synsets in D on average. Then, for each disambiguated instance $l \in L$, we retrieve all the candidate senses of its surface form in BabelNet and calculate a NASARI score N_s for each candidate sense. N_s is calculated as the cosine similarity between the centroid μ and its corresponding NASARI vector $NASARI(s)$:

$$N_s = Sim(\mu, NASARI(s)) \tag{3}$$

This score enables us to discard low-confidence disambiguated instances and correct the original disambiguation output from Babelfy in certain cases. Each $l \in L$ is re-tagged with the sense obtaining the highest NASARI score, provided that it exceeds 0.75:[15]

$$\hat{s} = \underset{s \in S_l}{\text{argmax}}\, N_s \tag{4}$$

where S_l is the set containing all the candidate senses for l.

In our running example (Example 1) Babelfy did not provide a high-confidence disambiguation for the word *king*, which was then incorrectly disambiguated using the MCS strategy. This error is corrected with the refinement step, as the chess-

[14] Throughout this step we represent each disambiguated instance as its corresponding synset in BabelNet.

[15] This threshold is set for the *refined* release of SensesDefs, which is intended to provide high-precision (rather than high-coverage) annotations.

related sense of *king* achieves higher semantic similarity with the disambiguated instances in $D \setminus L$, compared to its predominant sense. As shown in Fig. 1, the fact that we gathered definitions in different languages for the same concept proved essential in this disambiguation decision, as it provides a considerably larger context than the one given by a single definition.

4 SenseDefs: overview of the resource

By applying the methodology described in Sect. 3 on the whole set of textual definitions in BabelNet for all the available languages, we obtain a large multilingual corpus of disambiguated glosses: SenseDefs. SenseDefs is publicly available at the following website: http://lcl.uniroma1.it/sensedefs. We release two versions of the resource:

- *Full*. This high-coverage version provides sense annotations for all content words as provided by Babelfy after the context-rich disambiguation (see Sect. 3.2) and *before* the refinement step.
- *Refined*. The refined, high-precision version of SenseDefs, instead, *only* includes the most confident sense annotations as computed by the refinement step (see Sect. 3.3).

Some relevant statistics of SenseDefs are presented in Sect. 4.1, while Sect. 4.2 illustrates the format of the release.

4.1 Statistics

Table 1 shows some general statistics of the *full* and *refined* versions of SenseDefs, divided by resource. The output of the *full* version is a corpus of 38,820,114 disambiguated glosses, corresponding to 8,665,300 BabelNet synsets and covering 263 languages and 5 different resources (Wiktionary, WordNet, Wikidata, Wikipedia and OmegaWiki). It includes 249,544,708 sense annotations (6.4 annotations per definition on average). The refined version of the resource includes

Table 1 Number of definitions and annotations of the *full* and *refined* versions of SenseDefs

	# Glosses		# Annotations	
	Full	Refined	Full	Refined
Wikipedia	29,792,245	28,904,602	223,802,767	143,927,150
Wikidata	8,484,267	8,002,375	22,769,436	17,504,023
Wiktionary	281,756	187,755	1,384,127	693,597
OmegaWiki	115,828	106,994	744,496	415,631
WordNet	146,018	133,089	843,882	488,730
Total	38,820,114	37,334,815	249,544,708	163,029,131

fewer, but more reliable, sense annotations (see Sect. 5.1), and a slightly reduced number of glosses containing at least one sense annotation. Wikipedia is the resource with by far the largest number of definitions and sense annotations, including almost 30 million definitions and over 140 million sense annotations in both versions of the corpus. Additionally, Wikipedia also features textual definitions for the largest number of languages (over 200).

Statistics by language Figures 2 and 3 display the number of definitions and sense annotations, respectively, divided by language.[16] As expected, English provides the largest number of glosses and annotations (5.8M glosses and 37.9M sense annotations in the refined version), followed by German and French. Even though the majority of sense annotations overall concern resource-rich languages (i.e. those featuring the largest amounts of definitional knowledge), the language rankings in Figs. 2 and 3 do not coincide exactly: this suggests, on the one hand, that some languages (such as Vietnamese and Spanish, both with higher positions in Fig. 3 compared to Fig. 2) actually benefit from a cross-lingual disambiguation strategy; on the other hand, it also suggests that there is still room for improvement, especially for some other languages (such as Swedish or Russian) where the tendency is reversed and the number of annotations is lower compared to the amount of definitional knowledge available.

Table 2 shows the number of annotations divided by part-of-speech tag and disambiguation source. In particular, the full version obtained as output of Step 2 (Sect. 3.2) comprises two disambiguation sources: Babelfy and the MCS back-off (used for low-confidence annotations). The refined version, instead, removes the MCS back-off, either by discarding or correcting the annotation with NASARI (Sect. 3.3). Additionally, 17% of the sense annotations obtained by Babelfy without resorting to the MCS back-off are also corrected or discarded. Assuming the coverage of the full version to be 100%,[17] the coverage of our system after the refinement step is estimated to be 65.3%. As shown in Table 2, discarded annotations mostly consist of verbs, adjectives and adverbs, which are often harder to disambiguate as they are very frequently not directly related to the definiendum. In fact, the coverage figure on noun instances is estimated to be 73.9% after refinement.

4.2 Released format

SenseDefs is released in two different formats: a human- and machine-readable XML divided by language and resource (Sect. 4.2.1), and NIF (Sect. 4.2.2).

[16] Only the top 15 languages are displayed in the figures.

[17] There is no straightforward way to estimate the coverage of a disambiguation system automatically. In our first step using Babelfy, we provide disambiguated instances for all content words (including multi-word expressions) from BabelNet and also for overlapping mentions. Therefore, the output of our first step, even if it is not perfectly accurate, may be considered to have full coverage.

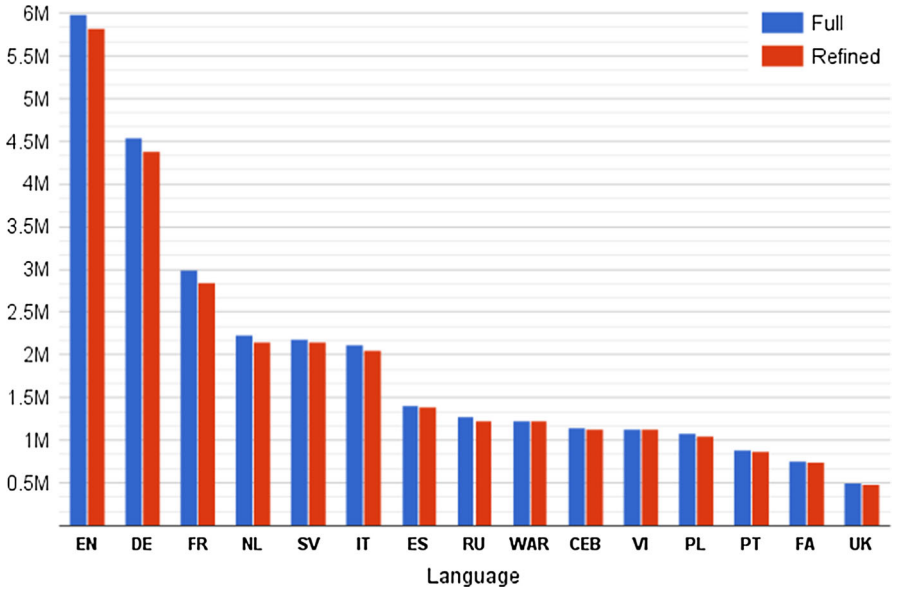

Fig. 2 Number of definitions by language (top 15 languages)

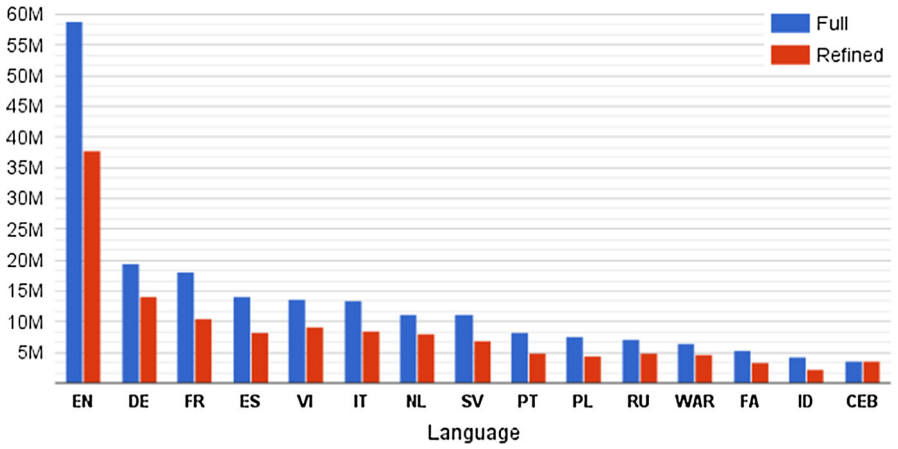

Fig. 3 Number of annotations by language (top 15 languages)

4.2.1 XML format

The format for each of the two versions of SENSEDEFS (*full* and *refined*) is almost identical: the corpus is first divided by resource (WordNet, Wikipedia, Wiktionary, Wikidata and OmegaWiki) and then divided by language within each resource.

The disambiguated glosses for each language and resource are stored in standard XML files. Figures 4 and 5 show a sample definition as displayed in the XML files

Table 2 Number of annotations by part-of-speech tag (*columns*) and by source (*rows*) before and after refinement

	All	Nouns	Verbs	Adjectives	Adverbs
FULL					
Babelfy	174,256,335	158,310,414	4,368,488	10,646,921	930,512
MCS	75,288,373	56,231,910	8 344,930	9,256,497	1,455,036
Total	249,544,708	214,542,324	12,713,418	19,903,418	2,385,548
REFINED					
Babelfy	144,637,032	140,111,921	1,326,947	3,064,416	133,748
NASARI	18,392,099	18,392,099	–	–	–
Total	163,029,131	158,504,020	1,326,947	3,064,416	133,748

of, respectively, the refined and full versions of SensesDefs. Each file contains a list of `definition` tags, with their respective `id`[18] as attribute. Then, each definition tag contains the original definition as plain text and its annotations. The `annotation` tag refers to the sense annotations provided as a result of our disambiguation process. Each annotation includes the BabelNet synset identifier and has four (or five) attributes (see Sect. 3 for more details about the attributes):

- `source` This indicates whether the disambiguation has been performed by Babelfy, the Most Common Sense (MCS) back-off (only in the full version of the corpus) or NASARI (only in the refined version of the corpus).
- `anchor` This corresponds to the exact surface-form match found within the definition.
- `bfScore` This corresponds to the Babelfy score.
- `coherenceScore` This corresponds to the coherence score.
- `nasariScore` This corresponds to the NASARI score (only in the refined version of the corpus).

4.2.2 NIF format

Recently the Linked Open Data community has made considerable efforts to extract and standardize structured knowledge from a wide range of corpora and linguistic resources, making them available on the Web by means of the RDF format (Chiarcos et al. 2011; Auer and Hellmann 2012; Ehrmann et al. 2014; Flati and Navigli 2014). In order to simplify the interoperability of linguistic resources, the NLP Interchange Format (NIF) was developed (Hellmann et al. 2013). NIF aims at easing the use of Linked Data among Natural Language Processing tools, language

[18] Identifiers depend on the resource, e.g. offsets in WordNet and page titles in Wikipedia.

```
<definition resource="WN" id="00166552n">
 <text>Interchanging the positions of the king and a rook</text>
  <annotations>
   <annotation source="MCS" anchor="king" bfScore="--" coherenceScore="--">
    bn:00049141n
   </annotation>
   <annotation source="BABELFY" anchor="rook" bfScore="0.8586" coherenceScore="0.4676">
    bn:00016571n
   </annotation>
   <annotation source="MCS" anchor="Interchanging" bfScore="--" coherenceScore="--">
    bn:00087802v
   </annotation>
   <annotation source="MCS" anchor="positions" bfScore="--" coherenceScore="--">
    bn:00062704n
   </annotation>
  </annotations>
</definition>
```

Fig. 4 Sample XML output for the definition of *castling* in WordNet from SENSEDEFS full

```
<definition resource="WN" id="00166552n">
 <text>Interchanging the positions of the king and a rook</text>
  <annotations>
   <annotation source="NASARI" anchor="king" bfScore="--" coherenceScore="--" nasariScore="0.8605">
    bn:00049147n
   </annotation>
   <annotation source="BABELFY" anchor="rook" bfScore="0.8586" coherenceScore="0.4676" nasariScore="0.8601">
    bn:00016571n
   </annotation>
  </annotations>
</definition>
```

Fig. 5 Sample XML output for the definition of *castling* in WordNet from SENSEDEFS refined

resources and annotations. Following this overarching goal, several resources have already been converted and made available on NIF format, contributing to the creation of the Linguistic Linked Open Data (Rizzo et al. 2012; Hellmann et al. 2012; Röder et al. 2014). In this paper we have transformed the English annotations of the refined version of SENSEDEFS into the NLP Interchange Format, following the guidelines provided by the hackathon organized at the Multilingual Linked Open Data for Enterprises Workshop (MLODE 2014).[19]

5 Evaluation

We evaluated SENSEDEFS both intrinsically (Sect. 5.1) and extrinsically on two Natural Language Processing tasks (Sect. 5.2).

5.1 Intrinsic evaluation

As intrinsic evaluation we carried out a thorough manual assessment of sense annotation quality in SENSEDEFS. In our previous study (Camacho-Collados et al. 2016a), we performed a manual evaluation for three languages (English, Italian and Spanish) employing three human judges. Each language was evaluated on a sample of 100 definitions, considering the input of a baseline (i.e. disambiguating definitions in isolation with Babelfy) and our *Full* and *Refined* versions of

[19] http://wwwusers.di.uniroma1.it/~flati/hackathon/index.html.

SensesDefs. In the three languages the context-rich disambiguation achieved better results than the baseline. More importantly, the refinement based on distributional similarity proved highly reliable, obtaining a precision over 90% on the three languages, without drastically decreasing the coverage. In this paper we have extended that intrinsic evaluation by performing two additional experiments. In the first experiment we extended the manual evaluation of Camacho-Collados et al. (2016a) by increasing the number of definitions, languages and annotators (Sect. 5.1.1). In the second experiment we performed a large-scale automatic evaluation where we compared our annotations against the manual disambiguation of WordNet glosses (Sect. 5.1.2).

5.1.1 Manual evaluation

We carried out an extensive evaluation of sense annotation quality in SensesDefs on four different languages: English, French, Italian and Spanish. To this end, we first randomly sampled 120 definitions for each language. Then, two annotators validated the sense annotations given by SensesDefs (both *Full* and *Refined*) and Babelfy. In contrast to the intrinsic evaluation of Camacho-Collados et al. (2016a), in this case we excluded those annotations coming from the MCS back-off, in order to assess the output explicitly provided by our disambiguation pipeline.

For each item in the sample, each annotator was shown the textual definition, the BabelNet entry for the definiendum, and every non-MCS sense annotation paired with the corresponding BabelNet entry. The annotator had to decide independently, for each sense annotation, whether it was correct (score of 1), or incorrect (score of 0). The disambiguation source (i.e. whether the annotation came from Babelfy in isolation, context-rich disambiguation or NASARI) was not shown. In some special cases where a certain sense annotation was acceptable but a more suitable synset was available, a score of 0.5 was allowed. One recurrent example of these indecisive annotations occurred on multi-word expressions: being designed as a high-coverage all-word disambiguation strategy, Babelfy can output disambiguation decisions over overlapping mentions when confronted with fragments of text having more than one acceptable disambiguation. For instance, the multi-word expression "*Commission of the European Union*" can be interpreted both as a single mention, referring to the specific BabelNet entity $\texttt{European\_Commission}^{1}_{n}$ (executive body of the European Union), and as two mentions, one ("*Commission*") referring to the BabelNet entry $\texttt{Parliamentary\_committee}^{1}_{n}$ (a subordinate deliberative assembly), and the other ("*European Union*") referring to the the BabelNet entry $\texttt{European\_Union}^{1}_{n}$ (the international organization of European countries). In all cases where one part of a certain multi-word expression was tagged with an acceptable meaning, but a more accurate annotation would have been the one associated with the whole multi-word expression, we allowed annotators to assign a score of 0.5 to valid annotations of nested mentions and a score of 1 only to the complete and correct multi-word annotation. Another controversial example of indecision is connected to semantic shifts due to Wikipedia redirections, which cause semantic annotations that are lexically acceptable but wrong from the point of

Table 3 Quality of the annotations of SENSEDEFS for English, Spanish, French and Italian

	#Annotations	Precision	Recall*	F1	IAA	
					ROA	κ
ENGLISH						
Babelfy	671	**84.3**	69.6	76.1	94.6	71.7
Full	714	80.0	70.2	74.8	94.2	70.1
Refined	**745**	83.1	**76.1**	**79.5**	95.3	71.9
SPANISH						
Babelfy	678	85.8	59.3	70.2	91.4	51.1
Full	**737**	82.6	62.1	70.9	92.4	66.2
Refined	725	**86.6**	**64.0**	**73.6**	95.1	63.3
FRENCH						
Babelfy	516	84.3	49.8	62.6	97.2	85.7
Full	568	81.3	52.8	64.0	96.7	86.4
Refined	**579**	**87.1**	**57.7**	**69.4**	95.1	65.8
ITALIAN						
Babelfy	540	**81.7**	53.5	64.7	94.5	74.3
Full	609	73.9	54.5	62.8	92.4	78.0
Refined	**618**	77.5	**58.1**	**66.4**	94.7	83.0

Bold numbers refer to best results overall in each language for each evaluation measure (the two last rows are inter-annotator agreements, no evaluation measures)

Recall (*) was computed assuming each content word in a sentence should be associated with a distinct sense. Inter-annotator agreement (IAA) was computed in terms of Relative Observed Agreement (ROA) and Cohen's kappa (κ). MCS annotations were not considered in this evaluation

view of semantic roles. For instance, the term *painter* inside Wikipedia redirects to the Wikipedia entry for `Painting` (*Graphic art consisting of an artistic composition made by applying paints to a surface*), while the term *Basketball player* redirects to the Wikipedia entry for `Basketball` (*Sport played by two teams of five players on a rectangular court*). These redirections are also exploited by Babelfy as acceptable disambiguation decisions (a policy that is often used in Entity Linking, especially in Wikipedia-specific settings) and, as such, they are also allowed a score of 0.5.

Once the annotations were completed, we calculated the Inter-Annotator Agreement (IAA) between the two annotators of each language by means of Relative Observed Agreement (ROA), calculated as the proportion of equal answers, and Cohen's kappa (Cohen 1968, κ). Finally, the two annotators in each language adjudicated the answers which were judged with opposite values. Table 3 shows the results of this manual evaluation. In the four languages, our refined version of the corpus achieved the best overall results, consistently with the results of the previous intrinsic evaluation (Camacho-Collados et al. 2016a). SENSEDEFS achieved over 80% precision in three of the four considered languages, both in its

full and refined versions. For Italian the precision dropped to 73.9 and 77.5%, respectively, probably due to its lower coverage in BabelNet. Finally, it is worth noting that, for all the examined languages, both the full and refined versions of SenseDefs provided more annotations than using the Babelfy baseline on isolated definitions.

5.1.2 Automatic evaluation: WordNet glosses

To complement the manual intrinsic evaluation, we performed an additional large-scale automatic evaluation. We compared the WordNet annotations given by SenseDefs[20] with the manually-crafted annotations of the disambiguated glosses from the Princeton Gloss Corpus.[21] Similarly to the previous manual evaluation, we included a baseline based on Babelfy disambiguating the definitions sentence-wise in isolation and using the pre-trained models[22] of the IMS (Zhong and Ng 2010, It Makes Sense) supervised disambiguation system. IMS uses a SVM classifier including features based on surrounding words and local collocations. As in our previous experiment, we did not consider the annotations for which the MCS back-off strategy was activated on any of the comparison systems. Finally, as baseline we include the results of WordNet first sense (i.e. MCS) for the annotations disambiguated by each system. The MCS baseline has been shown to be hard to beat, especially for knowledge-based systems (Raganato et al. 2017). However, this baseline, which is computed from a sense-annotated corpus, is only available for the English WordNet. Therefore, it is not possible to use this MCS baseline accurately for languages other than English, and resources other than WordNet for which sense-annotated data is not available or is very scarce.

Table 4 shows the accuracy results (computed as the number of automatic annotations corresponding to the manual annotations divided by the total number of overlapping annotations) of SenseDefs, Babelfy and IMS[23] on the Princeton Gloss Corpus. SenseDefs achieved an accuracy of 76.4%, both in its full and refined versions. Nevertheless, the refined version attained a larger coverage, disambiguating a larger amount of instances. This result is relatively high considering the nature of the corpus, consisting of short and concise definitions for which the context is clearly limited. In fact, even if not directly comparable, the best systems in standard WSD SemEval competitions (where full documents are given as context to disambiguate) tend to obtain considerably less accurate results (Edmonds and Cotton 2001; Snyder and Palmer 2004; Pradhan et al. 2007; Navigli et al. 2013;

[20] As explained in Sect. 3, our disambiguation pipeline annotates with BabelNet synsets, hence its coverage is larger than only WordNet. This implies that some annotations are not comparable to those inside the WordNet glosses.

[21] http://wordnet.princeton.edu/glosstag.shtml.

[22] Downloaded from http://www.comp.nus.edu.sg/~nlp/corpora.html#onemilwsd. We used the models from the One Million Sense-Tagged Instances as training corpus.

[23] Note that only IMS disambiguates all instances in the corpus. The reason why the recall of other systems is lower is twofold: first, IMS disambiguates all content words, unlike all other systems which use a confidence threshold; and second, it disambiguates all words with WordNet synsets, while in the other systems BabelNet is used as sense inventory (WordNet being a subset of BabelNet).

Table 4 Accuracy and number of compared WordNet annotations on the Princeton Gloss Corpus

	#WN annotations	Accuracy	MCS-Acc.
$\text{SENSEDEFS}_{\text{Full}}$	162 819 (59.0%)	**76.4**	66.1
$\text{SENSEDEFS}_{\text{Refined}}$	169 696 (61.5%)	**76.4**	65.2
Babelfy	130 236 (47.2%)	69.1	65.6
IMS	275 893 (100%)	56.1	55.2

Bold numbers refer to the best accuracy results overall

On the right the accuracy of the MCS baseline on the same sample

Moro and Navigli 2015). In fact, even though results are not directly comparable,[24] IMS achieved an accuracy which is considerably lower than our system's performance and also lower compared to its performance on standard benchmarks (Raganato et al. 2017). This result highlights the added difficulty of disambiguating definitions, as they do not provide enough context for an accurate disambiguation in isolation. Only our disambiguation pipeline, which does not make use of any sense-annotated data, proves reliable in this experiment, comfortably outperforming the MCS baseline on the same annotations.

5.2 Extrinsic evaluation

We also evaluated extrinsically the effectiveness of SENSEDEFS (both the *full* and *refined* versions of the resource) by making use of its sense annotations within two Natural Language Processing tasks.

The first experiment evaluated the full version of SENSEDEFS (before refinement) on Open Information Extraction (OIE) (Sect. 5.2.1). The experiment uses DEFIE (Delli Bovi et al. 2015), an OIE system designed to work on textual definitions. In its original implementation DEFIE used Babelfy to disambiguate definitions one-by-one before extracting relation instances. We modified that implementation and used the disambiguated glosses as obtained with our approach as input for the system, and then we compared the extractions with those obtained by the original implementation.

The second experiment, instead, evaluated the refined version of SENSEDEFS on the Sense Clustering task (Sect. 5.2.2). For this experiment we used the semantic representations of NASARI (see Sect. 3.3). In particular, we reconstructed the vectorial representations of NASARI by, (1) enriching the semantic network used in the original implementation with the refined sense annotations of SENSEDEFS, and (2) running again the NASARI pipeline to generate the vectors. We then evaluated these on the Sense Clustering task.

[24] Recall that our system annotates with BabelNet synsets and hence the set of disambiguation candidates is larger than IMS and the MCS baseline. This also makes the set of annotations differ with respect to IMS.

Table 5 Extractions of DEFIE on the evaluation sample

	# Glosses	# Triples	# Relations
DEFIE + GLOSSES	150	340	184
DEFIE	146	318	171

Table 6 Precision of DEFIE on the evaluation sample

	Relation	Relation instances
DEFIE + GLOSSES	0.872	0.780
DEFIE	0.865	0.770

5.2.1 Open information extraction

In this experiment we investigated the impact of our disambiguation approach on the definitional corpus used as input for the pipeline of DEFIE. The original OIE pipeline of the system takes as input an unstructured corpus of textual definitions, which are then preprocessed one-by-one to extract syntactic dependencies and disambiguate word senses and entity mentions. After this preprocessing stage, the algorithm constructs a syntactic-semantic graph representation for each definition, from which subject-verb-object triples (relation instances) are eventually extracted. As highlighted in Sect. 3.2, poor context of particularly short definitions may introduce disambiguation errors in the preprocessing stage, which then tend to propagate and reflect on the extraction of both relations and relation instances. To assess the quality of our disambiguation strategy as compared to the standard approach, we modified the implementation of DEFIE to consider our disambiguated instances instead of executing the original disambiguation step, and then we evaluated the results obtained at the end of the pipeline in terms of quality of relation and relation instances.

Experimental setup We first selected a random sample of 150 textual definitions from our disambiguated corpus (Sect. 4.1). We generated a baseline for the experiment by discarding all disambiguated instances from the sample, and treating the sample itself as an unstructured text of textual definitions which we used as input for DEFIE, letting the original pipeline of the system carry out the disambiguation step. Then we carried out the same procedure using, instead, the modified implementation for which our disambiguated instances are taken into account. In both cases, we ran the extraction algorithm of DEFIE and evaluated the output in terms of both relations and relation instances. Following Delli Bovi et al. (2015), we employed two human judges and performed the same evaluation procedure described therein over the set of distinct relations extracted from the sample, as well as the set of extracted relation instances.

Results Results reported in Tables 5 and 6 show a slight but consistent improvement resulting from our disambiguated glosses over both the number of extracted relations and triples and over the number of glosses with at least one extraction (Table 5), as well as over the estimated precision of such extractions (Table 6). Context-rich disambiguation of glosses across resources and languages

Table 7 Accuracy (Acc.) and F-measure (F1) percentages of different systems on the Wikipedia sense clustering datasets

Bold numbers refer to the best results overall in each dataset and for each evaluation measure (Acc. and F1)

	500-pair		SemEval	
	Acc.	F1	Acc.	F1
NASARI + SENSEDEFS	**86.0**	**74.8**	**88.1**	**64.7**
NASARI	81.6	65.4	85.7	57.4
SVM-monolingual	77.4	–	83.5	–
SVM-multilingual	84.4	–	85.5	–
Baseline	28.6	44.5	17.5	29.8

enabled the extraction of 6.5% additional instances from the sample (2.26 extractions on the average from each definition) and, at the same time, increased the estimated precision of relation and relation instances over the sample by $\sim$ 1%.

5.2.2 Sense clustering

Our second experiment focuses on the sense clustering task. Knowledge resources such as Wikipedia or WordNet suffer from the high granularity of their sense inventories. A meaningful clustering of senses within these sense inventories could help boost the performance in different applications (Hovy et al. 2013; Mancini et al. 2017; Pilehvar et al. 2017). In the following we explain how to deal with this issue in Wikipedia.

Our method for clustering senses in Wikipedia was based on the semantic representations of NASARI (Camacho-Collados et al. 2016b). We integrated the high-precision version of the network as an enrichment of the BabelNet semantic network, in order to improve the results of the state-of-the-art system based on the NASARI lexical vectors. NASARI uses Wikipedia ingoing links and the BabelNet taxonomy in the process of obtaining contextual information for a given concept. We simply enriched the BabelNet taxonomy with the refined version of the disambiguated glosses (see Sect. 3.3) of the target language. These disambiguated glosses contain synsets that are highly semantically connected with the definiendum, which makes them particularly suitable for enriching a semantic network. The rest of the pipeline for obtaining lexical semantic representations (i.e. lexical specificity applied to the contextual information) remained unchanged. By integrating the high-precision disambiguated glosses into the NASARI pipeline, we obtained a new set of vector representations for BabelNet synsets, increasing its initial coverage (4.4M synsets covered by the original NASARI, compared to 4.6M synsets covered by NASARI enriched with our disambiguated glosses).

Experimental setup We used the two sense clustering datasets constructed by Dandala et al. (2013). In these datasets sense clustering is viewed as a binary classification task. Given a pair of Wikipedia articles, the task consists of deciding whether they should be merged into a single cluster or not. The first dataset (*500-pair* henceforth) contains 500 pairs of Wikipedia articles, while the second dataset (*SemEval*) consists of 925 pairs coming from a set of highly ambiguous words taken from WSD Semeval competitions (Mihalcea 2007). We followed the original setting

of Camacho-Collados et al. (2016b) and clustered a pair of Wikipedia articles only when their similarity, computed by using the square-rooted Weighted Overlap comparison measure (Pilehvar et al. 2013), was above 0.5 (i.e. the middle point in the Weighted Overlap similarity scale).

Results Table 7 shows the accuracy and F1 results in the sense clustering task. As a comparison we included the Support Vector Machine classifier of Dandala et al. (2013), which exploits information from Wikipedia in English (*SVM-monolingual*) and four different languages (*SVM-multilingual*). As a simple baseline we additionally included a system which clusters all pairs. Finally, we report the results of the original NASARI English lexical vectors (NASARI)[25] and the NASARI-based vectors obtained from the enriched BabelNet semantic network (NASARI + SENSEDEFS). As shown in Table 7, the enrichment produced by our glosses proved to be highly beneficial, significantly improving on the original results obtained by NASARI. Moreover, NASARI + SENSEDEFS obtained the best performance overall, outperforming the SVM-based systems of Dandala et al. (2013) in terms of accuracy in both datasets.

6 Related work

Word Sense Disambiguation is a long-standing task in Natural Language Processing (NLP), lying at the very core of language understanding (Navigli 2009). However, the lack of sense-annotated data is slowing down progress in the field, as the largest manually sense-annotated dataset (for WordNet) dates back to the nineties: Miller et al (1993, SemCor). This is mainly due to the expensive manual effort required to annotate large corpora. In order to overcome this gap, several recent studies have proposed different automatic approaches to obtain reliable and large-scale sense-annotated data (Pilehvar and Navigli 2014; Taghipour and Ng 2015; Raganato et al. 2016; Pasini and Navigli 2017), which have been shown to improve the performance of supervised WSD systems (Raganato et al. 2017).

In particular, disambiguating definitions has attracted a considerable amount of interest. To date, WordNet has definitely been the most popular and the most exploited resource among those that include textual definitions. In fact, WordNet glosses have still been used successfully in recent work (Khan et al. 2013; Chen et al. 2015). A first attempt to disambiguate WordNet glosses automatically was proposed as part of the eXtended WordNet project[26] (Novischi 2002). However, this attempt's estimated coverage did not reach 6% of the total number of sense-annotated instances. Moldovan and Novischi (2004) proposed an alternative disambiguation approach, specifically targeted at the WordNet sense inventory and based on a supervised model trained on SemCor (Miller et al. 1993). In general, the drawback of using supervised models arises from the so-called *knowledge-acquisition bottleneck*, a problem that becomes particularly vexed when such models are applied to larger inventories, due to the vast amount of annotated data

[25] Downloaded from http://lcl.uniroma1.it/nasari/.

[26] http://www.hlt.utdallas.edu/~xwn/.

they normally require. Another disambiguation task focused on WordNet glosses was presented as part of the Senseval-3 workshop (Litkowski 2004). However, the best reported system obtained precision and recall figures below 70%, which is arguably not enough to provide high-quality sense-annotated data for current state-of-the-art NLP systems.

In addition to annotation reliability, another issue that arises when producing a corpus of textual definitions is coverage. In fact, reliable corpora of sense-annotated definitions produced to date, such as the Princeton WordNet Gloss Corpus,[27] have usually been obtained employing human annotators. The Princeton corpus of WordNet disambiguated glosses has already been shown to be successful as part of the pipeline in semantic similarity (Pilehvar et al. 2013), domain labeling (González et al. 2012) and Word Sense Disambiguation (Agirre and Soroa 2009; Camacho-Collados et al. 2015b) systems. However, as new encyclopedic knowledge about the world is constantly being harvested, keeping up using only human annotation is becoming an increasingly expensive endeavor. With a view to tackling this problem, a great deal of research has recently focused on the automatic extraction of definitions from unstructured text (Navigli and Velardi 2010; Benedictis et al. 2013; Espinosa-Anke and Saggion 2014; Dalvi et al. 2015). At the same time, the prominent role of collaborative resources (Hovy et al. 2013) has created a convenient development ground for NLP systems based on encyclopedic definitional knowledge. By bridging the gap between lexicographic and encyclopedic knowledge, BabelNet (Navigli and Ponzetto 2012) is a key milestone in this respect. BabelNet includes, among others, Wikipedia as an additional source of encyclopedic knowledge, thus enabling the application of Entity Linking techniques.[28] Nevertheless, extending the manual annotation of definitions to such larger and up-to-date knowledge repositories is clearly not feasible. First of all, the number of items to disambiguate is massive; moreover, as the number of concepts and named entities increases, annotators would have to deal with the added difficulty of selecting context-appropriate synsets from an extremely large sense inventory. In fact, WordNet 3.0 comprises 117,659 synsets and a definition for each synset, while BabelNet 3.0 covers 13,801,844 synsets with a total of 40,328,194 definitions.

With the aim of overcoming this shortfall, in this paper we propose an automatic disambiguation approach which leverages multilinguality and cross-resource information along with a state-of-the-art graph-based WSD and Entity Linking system (Moro et al. 2014) and a distributional representation of concepts and entities (Camacho-Collados et al. 2015a). By exploiting at best all these components, we are able to produce a large-scale high-quality corpus of glosses, SENSEDEFS, automatically disambiguated with BabelNet synsets.

[27] http://wordnet.princeton.edu/glosstag.shtml.

[28] See Usbeck et al. (2015) and Ling et al. (2015) for an overview and comparison of entity linking systems.

7 Conclusion

In this paper we presented SENSEDEFS, a large-scale multilingual corpus of disambiguated textual definitions (or glosses). We obtained high-quality sense annotations with a disambiguation pipeline designed to exploit cross-resource and cross-language complementarities of multiple textual definitions associated with a given definiendum. By leveraging the structure of a wide-coverage semantic network and sense inventory like BabelNet, we obtained a corpus of textual definitions coming from multiple sources and multiple languages, fully disambiguated with BabelNet synsets. SENSEDEFS, to the best of our knowledge, is the largest available corpus of its kind. Moreover, the choice of BabelNet as sense inventory not only provides wide-coverage sense annotations of both a lexicographic and encyclopedic nature. Indeed, since BabelNet is a merger of various different resources, including WordNet and Wikipedia, these annotations are also expandable to any of these resources and can be easily converted via BabelNet's inter-resource mappings.

SENSEDEFS is based on the very large and heterogeneous corpus of textual definitions provided by BabelNet. After collecting all the definitions of a given concept or entity into a single multilingual text, our pipeline carries out disambiguation in two subsequent stages. In the first stage, we leverage a state-of-the-art multilingual disambiguation system, Babelfy (Moro et al. 2014), which is designed to exploit at best a multiple-language setting. Using Babelfy, we obtain an initial set of sense annotations for all the available languages of the target corpus. These initial sense annotations are then refined in the second stage, by integrating a module based on NASARI (Camacho-Collados et al. 2016b) and distributional similarity targeted to identify a subset of sense annotations disambiguated with high-confidence. This refined version of the corpus was proven very reliable, with precision and coverage figures over 80 and 60%, respectively.

We release to the research community two versions of SENSEDEFS: a full version comprising all the sense annotations obtained with Babelfy in the first stage, and a refined version including only the high-confidence annotations identified through distributional similarity. Both versions additionally include a set of confidence scores which can be taken into account by users for tuning them to their needs. The refined version is especially suitable for high-precision applications, where having a disambiguation error as low as possible is the foremost requirement. Moreover, since high-precision sense annotations are those that are most closely connected to the definiendum, they can also be used to enrich a semantic network (or to build a semantic network on its own). The full version is, instead, targeted at applications requiring high coverage, where extracting as much information as possible is key, even at the cost of lower-confidence disambiguation decisions. In knowledge acquisition and extraction, for instance, it could be just as important to discover semantic relations between the definiendum and a concept or entity that is not part of the same domain of knowledge.

We evaluated SENSEDEFS extensively, with both intrinsic and extrinsic experiments. We assessed sense annotation quality intrinsically on four different

languages, showing the reliability of our system in comparison to previous approaches and to an off-the-shelf state-of-the-art disambiguation system. Finally, we also carried out an extrinsic evaluation where we showed two possible applications of our resource in both its full and refined versions, namely Open Information Extraction (a high-coverage setting) and Sense Clustering (a high-precision setting). In both cases, our sense annotations led to performance improvement and showed the flexibility of SENSEDEFS across different Natural Language Processing tasks. As future work we plan to exploit this corpus in two main directions: first, as a sense-annotated training corpus for supervised Word Sense Disambiguation and Entity Linking; second, in applications such as Taxonomy Learning for which definitional knowledge has proved beneficial. Moreover, we plan to investigate the effectiveness of our cross-lingual disambiguation strategy outside definitional knowledge, e.g. on general sentence-aligned parallel corpora as already proved effective in Delli Bovi et al. (2017).

Acknowledgements The authors gratefully acknowledge the support of the ERC Consolidator Grant MOUSSE No. 726487. Jose Camacho-Collados was supported by a Google Ph.D. Fellowship in Natural Language Processing. We would like to thank Ilenia Giambruno for helping us with the intrinsic evaluation. Finally, we would also like to thank Jim McManus for his comments on the manuscript.

References

Agirre, E., & Soroa, A. (2009). Personalizing PageRank for Word Sense Disambiguation. In *Proceedings of EACL* (pp. 33–41).

Auer, S., & Hellmann, S. (2012). The web of data: Decentralized, collaborative, interlinked and interoperable. In *Proceedings of the 8th international conference on language resources and evaluation (LREC-2012)*.

Banerjee, S., & Pedersen, T. (2002). An adapted Lesk algorithm for Word Sense Disambiguation using WordNet. In *Proceedings of the 3rd international conference on computational linguistics and intelligent text processing, Mexico City, Mexico, CICLing'02* (pp. 136–145).

Basile, P., Caputo, A., & Semeraro, G. (2014). An enhanced lesk word sense disambiguation algorithm through a distributional semantic model. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers, Dublin, Ireland* (pp. 1591–1600).

Benedictis, F. D., Faralli, S., & Navigli, R. (2013). GlossBoot: Bootstrapping multilingual domain glossaries from the web. In *Proceedings of ACL* (pp. 528–538).

Camacho-Collados, J., & Navigli, R. (2017). BabelDomains: Large-scale domain labeling of lexical resources. In *Proceedings of EACL (2), Valencia, Spain* (pp. 223–228).

Camacho-Collados, J., Pilehvar, M. T., & Navigli, R. (2015a). NASARI: A novel approach to a semantically-aware representation of items. In *Proceedings of NAACL* (pp. 567–577).

Camacho-Collados, J., Pilehvar, M. T., & Navigli, R. (2015b). A unified multilingual semantic representation of concepts. In *Proceedings of ACL, Beijing, China* (pp. 741–751).

Camacho-Collados, J., Bovi, C. D., Raganato, A., & Navigli, R. (2016a). A large-scale multilingual disambiguation of glosses. In *Proceedings of LREC, Portoroz, Slovenia* (pp. 1701–1708).

Camacho-Collados, J., Pilehvar, M. T., & Navigli, R. (2016b). Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence*, *240*, 36–64.

Chen, D., Fisch, A., Weston, J., & Bordes, A. (2017). Reading wikipedia to answer open-domain questions. In *Proceedings of ACL* (pp. 1870–1879).

Chen, T., Xu, R., He, Y., & Wang, X. (2015). Improving distributed representation of word sense via wordnet gloss composition and context clustering. In *Proceedings of ACL* (pp. 15–20).

Chen, X., Liu, Z., & Sun, M. (2014). A unified model for word sense representation and disambiguation. In *Proceedings of EMNLP, Doha, Qatar* (pp. 1025–1035).

Chiarcos, C., Hellmann, S., & Nordhoff, S. (2011). Towards a linguistic linked open data cloud: The open linguistics working group. *TAL*, *52*(3), 245–275.

Cocos, A., Apidianaki, M., & Callison-Burch, C. (2017). Mapping the paraphrase database to wordnet. In *Proceedings of the 6th joint conference on lexical and computational semantics (*SEM 2017)* (pp. 84–90).

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213.

Cucerzan, S. (2007). Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of EMNLP-CoNLL* (pp. 708–716).

Dalvi, B., Minkov, E., Talukdar, P. P., & Cohen, W. W. (2015). Automatic gloss finding for a knowledge base using ontological constraints. In *Proceedings of WSDM* (pp. 369–378).

Dandala, B., Hokamp, C., Mihalcea, R., & Bunescu, R. C. (2013). Sense clustering using Wikipedia. In *Proceedings of recent advances in natural language processing, Hissar, Bulgaria* (pp. 164–171).

Delli Bovi, C., Telesca, L., & Navigli, R. (2015). Large-scale information extraction from textual definitions through deep syntactic and semantic analysis. In *Transactions of the Association for Computational Linguistics (TACL) 3*.

Delli Bovi, C., Camacho-Collados, J., Raganato, A., & Navigli, R. (2017). Eurosense: Automatic harvesting of multilingual sense annotations from parallel text. In *Proceedings of ACL (2)* (pp. 594–600).

Edmonds, P., & Cotton, S. (2001). Senseval-2: Overview. In *Proceedings of the 2nd international workshop on evaluating word sense disambiguation systems, Toulouse, France* (pp. 1–6).

Ehrmann, M., Cecconi, F., Vannella, D., McCrae, J. P., Cimiano, P., & Navigli, R. (2014). Representing multilingual data as linked data: The case of babelnet 2.0. In *LREC* (pp. 401–408).

Espinosa-Anke, L., & Saggion, H. (2014). Applying dependency relations to definition extraction. *Natural Language Processing and Information Systems*, *8455*, 63–74.

Espinosa-Anke, L., Camacho-Collados, J., Delli Bovi, C., & Saggion, H. (2016a). Supervised distributional hypernym discovery via domain adaptation. In *Proceedings of EMNLP* (pp. 424–435).

Espinosa-Anke, L., Saggion, H., Ronzano, F., & Navigli, R. (2016b). ExTaSem! extending, taxonomizing and semantifying domain terminologies. In *Proceedings of the 30th conference on artificial intelligence (AAAI'16)*.

Faralli, S., & Navigli, R. (2012). A new minimally-supervised framework for domain word sense disambiguation. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning, association for computational linguistics* (pp. 1411–1422).

Fernandez-Ordonez, E., Mihalcea, R., & Hassan, S. (2012). Unsupervised word sense disambiguation with multilingual representations. In *LREC* (pp. 847–851).

Flati, T., & Navigli, R. (2014). Three birds (in the llod cloud) with one stone: Babelnet, babelfy and the wikipedia bitaxonomy. In *Proceedings of SEMANTiCS2014*.

Flati, T., Vannella, D., Pasini, T., & Navigli, R. (2016). MultiWiBi: The multilingual Wikipedia bitaxonomy project. *Artificial Intelligence*, *241*, 66–102.

Franco-Salvador, M., Rosso, P., & Montes-y Gómez, M. (2016). A systematic study of knowledge graph analysis for cross-language plagiarism detection. *Information Processing & Management*, *52*, 550–570.

Gabrilovich, E., & Markovitch, S. (2007). Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of IJCAI* (pp. 1606–1611).

Gale, W. A., Church, K., & Yarowsky, D. (1992). A method for disambiguating word senses in a corpus. *Computers and the Humanities*, *26*, 415–439.

González, A., Rigau, G., & Castillo, M. (2012). A graph-based method to improve Wordnet domains. In *Proceedings of 13th international conference on intelligent text processing and computational linguistics (CICLING), New Delhi, India* (pp. 17–28).

Hellmann, S., Stadler, C., & Lehmann, J. (2012). The german dbpedia: A sense repository for linking entities. In *Linked data in linguistics* (pp. 181–190). Springer.

Hellmann, S., Lehmann, J., Auer, S., & Brümmer, M. (2013). Integrating nlp using linked data. In *International semantic web conference* (pp. 98–113). Springer.

Hill, F., Cho, K., Korhonen, A., & Bengio, Y. (2015). Learning to understand phrases by embedding the dictionary. arXiv preprint arXiv:150400548.

Hovy, E. H., Navigli, R., & Ponzetto, S. P. (2013). Collaboratively built semi-structured content and artificial intelligence: The story so far. *Artificial Intelligence*, *194*, 2–27.

Khan, M. F., Khan, A., & Khan, K. (2013). Efficient word sense disambiguation technique for sentence level sentiment classification of online reviews. *Science International (Lahore)*, *25*, 937–943.

Lafon, P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *Mots*, *1*, 127–165.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual conference on systems documentation, Toronto, Ontario, Canada* (pp. 24–26).

Lieto, A., Mensa, E., & Radicioni, D. P. (2016). A resource-driven approach for anchoring linguistic resources to conceptual spaces. In: *AI* IA 2016 advances in artificial intelligence* (pp. 435–449). Springer.

Ling, X., Singh, S., & Weld, D. S. (2015). Design challenges for entity linking. *Transactions of the Association for Computational Linguistics*, *3*, 315–328.

Litkowski, K. C. (2004). Senseval-3 task: Word-sense disambiguation of wordnet glosses. In *Proceedings of SENSEVAL-3 workshop on sense evaluation, in the 42th annual meeting of the association for computational linguistics (ACL 2004), Citeseer*.

Mancini, M., Camacho-Collados, J., Iacobacci, I., & Navigli, R. (2017). Embedding words and senses together via joint knowledge-enhanced training. In *Proceedings of CoNLL, Vancouver, Canada* (pp. 100–111).

Mihalcea, R. (2007). Using Wikipedia for automatic Word Sense Disambiguation. In *Proceedings of NAACL-HLT-07, Rochester, NY* (pp. 196–203).

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. CoRR abs/1301.3781, http://arxiv.org/abs/1301.3781.

Miller, G. A., Beckwith, R., Fellbaum, C. D., Gross, D., & Miller, K. (1990). WordNet: An online lexical database. *International Journal of Lexicography*, *3*(4), 235–244.

Miller, G.A., Leacock, C., Tengi, R., Bunker, R. (1993). A semantic concordance. In *Proceedings of the 3rd DARPA workshop on human language technology* (pp. 303–308).

Moldovan, D., & Novischi, A. (2004). Word sense disambiguation of wordnet glosses. *Computer Speech & Language*, *18*(3), 301–317.

Moro, A., & Navigli, R. (2015). Semeval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. In *Proceedings of SemEval-2015*.

Moro, A., Raganato, A., & Navigli, R. (2014). Entity linking meets Word Sense Disambiguation: A unified approach. *Transactions of the Association for Computational Linguistics (TACL)*, *2*, 231–244.

Navigli, R. (2009). Word Sense Disambiguation: A survey. *ACM Computing Surveys*, *41*(2), 1–69.

Navigli, R. (2012). A quick tour of word sense disambiguation, induction and related approaches. In *SOFSEM 2012: Theory and practice of computer science* (pp. 115–129). Springer.

Navigli, R., & Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, *193*, 217–250.

Navigli, R., & Velardi, P. (2005). Structural semantic interconnections: A knowledge-based approach to Word Sense Disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(7), 1075–1088.

Navigli, R., & Velardi, P. (2010). Learning Word-Class Lattices for definition and hypernym extraction. In *Proceedings of ACL 2010, Uppsala, Sweden* (pp. 1318–1327).

Navigli, R., Jurgens, D., & Vannella, D. (2013). SemEval-2013 Task 12: Multilingual Word Sense Disambiguation. *Proceedings of SemEval*, *2013*, 222–231.

Nivre, J., de Marneffe, M. C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., & Zeman, D. (2016). Universal dependencies v1: A

multilingual treebank collection. In *Proceedings of the 10th international conference on language resources and evaluation (LREC 2016)*. European Language Resources Association (ELRA).

Novischi, A. (2002). Accurate semantic annotations via pattern matching. In *FLAIRS conference* (pp. 375–379).

Pasini, T., & Navigli, R. (2017). Train-o-matic: Large-scale supervised word sense disambiguation inmultiple languages without manual training data. In *Proceedings of empirical methods in natural language processing, Copenhagen, Denmark*.

Pilehvar, M. T., & Navigli, R. (2014). A large-scale pseudoword-based evaluation framework for state-of-the-art Word Sense Disambiguation. *Computational Linguistics*, *40*(4), 837–881.

Pilehvar, M. T., Jurgens, D., & Navigli, R. (2013). Align, disambiguate and walk: A unified approach for measuring semantic similarity. In *Proceedings of the 51st annual meeting of the association for computational linguistics, Sofia, Bulgaria* (pp. 1341–1351).

Pilehvar, M. T., Camacho-Collados, J., Navigli, R., & Collier, N. (2017). Towards a seamless integration of word senses into downstream nlp applications. In *Proceedings of ACL* (pp. 1857–1869).

Pradhan, S., Loper, E., Dligach, D., & Palmer, M. (2007). SemEval-2007 task-17: English lexical sample, SRL and all words. In *Proceedings of SemEval* (pp. 87–92).

Raganato, A., Delli Bovi, C., & Navigli, R. (2016). Automatic construction and evaluation of a large semantically enriched Wikipedia. In *Proceedings of IJCAI, New York City, NY, USA* (pp. 2894–2900).

Raganato, A., Camacho-Collados, J., & Navigli, R. (2017). Word Sense Disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of EACL, Valencia, Spain* (pp. 99–110).

Richardson, S. D., Dolan, W. B., & Vanderwende, L. (1998). MindNet: Acquiring and structuring semantic information from text. In *Proceedings of ACL* (pp. 1098–1102).

Rizzo, G., Troncy, R., Hellmann, S., & Bruemmer, M. (2012). NERD meets NIF: Lifting NLP Extraction Results to the Linked Data Cloud.. In *LDOW 937*.

Röder, M., Usbeck, R., Hellmann, S., Gerber, D., & Both, A. (2014). N3-a collection of datasets for named entity recognition and disambiguation in the nlp interchange format. In *9th LREC*.

Shalaby, W., & Zadrozny, W. (2015). Measuring semantic relatedness using mined semantic analysis. arXiv preprint arXiv:151203465.

Snyder, B., & Palmer, M. (2004). The English all-words task. In *Proceedings of the 3rd international workshop on the evaluation of systems for the semantic analysis of text (SENSEVAL-3), Barcelona, Spain, Barcelona, Spain* (pp. 41–43).

Taghipour, K., & Ng, H. T. (2015). One million sense-tagged instances for word sense disambiguation and induction. *CoNLL*, *2015*, 338.

Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 conference of the North American chapter of the association for computational linguistics on human language technology* (Vol. 1, pp. 173–180), Association for Computational Linguistics.

Tripodi, R., & Pelillo, M. (2017). A game-theoretic approach to word sense disambiguation. *Computational Linguistics*, *43*(1), 31–70.

Usbeck, R., Röder, M., Ngonga Ngomo, A. C., Baron, C., Both, A., Brümmer, M., Ceccarelli, D., Cornolti, M., Cherix, D., & Eickmann, B., et al. (2015). Gerbil: General entity annotator benchmarking framework. In *Proceedings of the 24th international conference on World Wide Web* (pp. 1133–1143). International World Wide Web Conferences Steering Committee.

Velardi, P., Faralli, S., & Navigli, R. (2013). OntoLearn reloaded: A graph-based algorithm for taxonomy induction. *Computational Linguistics*, *39*(3), 665–707.

Vrandečić, D. (2012). Wikidata: A new platform for collaborative data collection. In *Proceedings of WWW* (pp. 1063–1064).

Wu, F., & Weld, D. S. (2010). Open information extraction using Wikipedia. In *Proceedings of ACL* (pp. 118–127).

Young, J., Basile, V., Kunze, L., Cabrio, E., & Hawes, N. (2016). Towards lifelong object learning by integrating situated robot perception and semantic web mining. In *Proceedings of the European conference on artificial intelligence conference, The Hague, Netherland* (pp. 1458–1466).

Zhong, Z., & Ng, H. T. (2010). It makes sense: A wide-coverage Word Sense Disambiguation system for free text. In *Proceedings of the ACL system demonstrations* (pp. 78–83).